J. American Voice I/O Society, 1986, Volume 3 pp 39-67

ISSUES IN THE EVALUATION OF SPEECH RECOGNITION SYSTEMS

M. M. Taylor[1]
DCIEM, Box 2000, Downsview, Ontario, Canada M3M 3B9

Abstract

Many factors come into play when Automatic Speech Recogition (ASR) devices are to be evaluated. Different kinds of information are required by designers, by manufacturers, and by potential purchasers. Some applications may require very accurate recognition of a few selected and trained talkers, whereas others require reasonably accurate recognition of untrained self-selected talkers. Error percentages are dependent on the test vocabulary, and may not even be monotonic with overall recognizer capacity. A measure called "Effective Vocabulary Capability" EVC is introduced. The EVC is conceptually independent of the vocabulary used to test the machine. The EVC of a device is the maximum size of vocabulary that the device could recognize with a tolerably small error rate, if the words in the vocabulary were chosen for maximum distinctiveness. EVC may vary across talkers, within a talker across training, or across environmental conditions. On initial exposure to an ASR device, some talkers may achieve a high EVC, whereas most obtain only a low EVC; such a machine is talker-specific. If only a few talkers fail to achieve a high EVC, the machine is talker-independent. The paper does not consider methods of determining the EVC for a particular talker on a particular machine.

---

[1] A longer version of this paper, containing a proposed method of computing EVC, was originally prepared in 1980 at the request of and with the assistance of NATO AC243 (Panel III/RSG-10) Research Study Group on Automatic Speech Processing, which has now requested that this shorter version be submitted for publication to a wider audience. Many members of RSG-10 have provided helpful comments on various drafts; special acknowledgement is due to the contributions of L. C. W. Pols (Netherlands), J. S. Bridle (United Kingdom), and R. K. Moore (United Kingdom). None of these should be held accountable for the undoubted errors and omissions. Detailed methods and experimental results for a rapid method of computing EVC will be presented in forthcoming papers.

Introduction

By 1980, Automatic Speech Recognition (ASR) devices were commercially available at prices ranging from $259 to $67,000 (Kaplan, 1980). These systems all claimed accuracies of over 95% on their test vocabularies. With such comparable performance claims and wide price range, how can a potential user assess the comparative value of ASR devices in his application?

In performance claims, little is usually said about such things as what constituted the test vocabularies, how the vocabularies were spoken, what proportion of talkers could achieve such high scores, and so forth. Even when such data are available, the differing test conditions make comparison of ASR devices very difficult. Is a score of 99% correct on a 200-word vocabulary enunciated carefully by a phonetically trained talker better than 97% on a list of connected digits spoken by several untrained talkers selected at random off the street? There are many such issues in specifying the performance of an ASR device. In some applications one aspect may be important, in others, a quite different aspect may dominate. For example, a public-access catalogue-order system may require moderate accuracy for digits and a few keywords over a large proportion of the population, whereas a complex control system may require that a few talkers can be trained to obtain a high accuracy over the same kind of vocabulary. The ASR device that is best for one application is not necessarily best for another.

A wider issue in the use of ASR devices is the question of whether ASR is appropriate at all in a given application. No ASR device could substitute for the keyboard of a piano, whereas an accurate and rapid one can serve as a good substitute for a keyboard in an inspection task on an assembly line, as, for example, in the Volkswagen plant at Wolfsburg, Germany. In some applications, the value of an ASR device will depend on its recognition performance, but it is beginning to be understood that the effective use of ASR devices in real applications requires a careful ergonomic study of the whole system. Evaluating the recognition performance of the ASR device is only a small part of the total study.

This paper does not address the ergonomic questions of ASR. Rather, it considers the evaluation of ASR devices in applications for which they are considered to be well suited. It attempts to provide methods of answering the question: "Which ASR device should I buy for this application", and not "Should I use ASR for this input?". It is aimed primarily at consultants or others who will make technical evaluations of ASR devices in real applications, and to a lesser extent at ASR researchers and end-users of ASR. The primary interest of RSG-10 is in ASR devices suitable for NATO military uses. Most of the issues apply more generally, however, and the discussion is cast largely in commercial rather than military terms.

Requirements for evaluation

Application-oriented Testing

Different people have different motives for evaluating ASR systems. The ASR researcher who is trying to build a better or cheaper ASR device needs to know just where the device is weakest, and where a small investment of effort will lead to a big improvement in performance. The manufacturer would like to have a standardized set of tests against which to try the device, so that the results can be used in advertising it to potential users. For the manufacturer, the results should suggest how the device will perform in a range of reasonably designed applications, but should be specific for none. The application user wants to know how well all the competing ASR devices will perform in the particular proposed application. At present the only way to do this is to buy samples of all the competitors, and test them under realistic conditions. We will illustrate the dilemma of application-oriented evaluation of ASR devices.

Evaluating systems for two ASR applications

The first question to be asked in designing any test is: "How is the application area defined?". Only by answering this question can appropriate sampling strategies and tests be developed. As an example, let us consider ASR systems for two applications:

1. The first is for a public access catalogue order system. The device must accept a small restricted query language as well as connected digit sequences, and must be part of a voice dialogue system working over a telephone line.

2. The second has the same requirements on task language, but is part of a control system for a complex processing plant, and little or no dialogue feedback is expected.

Requirements for the two systems

Both ASR systems are supposed to accept the same vocabulary, so one might think that comparative testing would be easy. The ASR device with highest accuracy should be chosen for both applications. Unfortunately, there is more to it than that. The applications have four major areas of difference:

- User population: The catalogue application must be open to the public, and therefore must tolerate a variety of dialects and speaking styles. The control application involves only a small set of skilled operators, who could be selected for high accuracy with the ASR device, and who could be trained to speak in a rigid style if the device required it.
- Dialogue: The catalogue application expects there to be a dialogue between the user and the system. This dialogue can be used to control the responses of the user and to limit the possibilities at any one moment. It can further be used to clear up ambiguous input from the user (*Please spell that?*), and to ensure that the correct catalogue item was requested (*You are ordering a brown pair of shoes. Is this correct?*). The control application has time-critical functions and does not expect the use of much dialogue, since the dialogue both takes time and distracts the operator from proper performance of other control functions. Hence, its input must ordinarily be accurate and unambiguous on first entry.
- Redundancy of input: In the catalogue application, the item numbering scheme can be chosen to incorporate redundancy and error-correcting coding. This is probably a good thing to do in a catalogue application even without the possibility of voice ordering, since it provides the assurance of transaction accuracy in the face of input error. Such assurance is almost mandatory in an ASR system that must deal with the public, but it implies also that the design of the ASR system can trade high accuracy with a few talkers for robustness in the face of talker variation. In the control application, on the other hand, redundancy cannot be guaranteed in the numerical part of the input, since almost any setting may be selected for a controlled variable. Hence, high input accuracy per digit is required.
- Speech quality: The catalogue application expects its input to be over the public telephone system, whereas the operator of the control system can wear a good noise-cancelling microphone. These two media provide vastly different voice quality, and two ASR systems may well have different robustness against noise.

In the light of these differences in application, testing the machines becomes an exercise in correct statistical generalization.

Differences in testing for the two applications

For the first application, it is necessary that the machine work with "almost everybody". Not only must there be many different talkers, to ensure that fewer than (say) 1% of the population will be excluded from using the final system, but also these talkers must be selected from all the different dialect groups who may be expected to want to order from the catalogue. If, for example, the system were in the USA,

and it consistently rejected black talkers, it could cause economic catastrophe for the company that used it. Hence, there must be enough talkers in the test population to ensure that the reject proportion is acceptably low in each individual dialect group.

For the control application, it is required only that a few highly selected talkers be trained to talk to the machine. Hence, the test need only use a few talkers, enough to ensure that the reject rate will be less than (say) 20% among the population from whom system operators are chosen. These rejects, moreover, are rejected only after attempts to train them to the machine and to adapt the machine to them have failed. On the other hand, rejection may occur when the error rate on the test vocabulary exceeds 0.1%, whereas for the catalogue application, an error rate of 2.0% may be acceptable.

To judge the magnitude of the statistical differences between tests that are appropriate for the two applications, consider the easiest test of an acceptable device: whether it attains a sufficiently low error rate on a prescribed vocabulary.

Let us suppose that we had an ASR device that was (unknown to us) error-free, and we wished to test whether it was sufficiently accurate for either the catalogue or the control-system application. For the catalogue application, we need to know whether it rejects (has a greater than 2% error rate for) more than 1% of any major dialect group. To test this question with 95% confidence, we would need to sample at least 150 utterances [2] from at least 300 talkers of each dialect group. For a similar test of the control application (20% rejects at 0.1% error rate), we would need only 14 talkers, but would need 3000 utterances from each talker. If an acceptable device made some errors, as real devices do, the numbers of talkers and the amount of speech from each would need to be larger before we could determine that it was in fact acceptable.

The statistical requirements of the two applications are very different. The catalogue system requires that thousands of people be tested, but with relatively little speech from each talker, whereas the control system requires only that some few people be found who can achieve high accuracy, and that large amounts of speech be tested from those people.

---

[2] The tester wishes to ascertain that the device makes less than 2% errors, with 95% confidence, but (by assumption) it makes actually makes no errors. A machine that made 2% error would have a 5% chance of producing no errors only after 149 trials ($0.98^{149} \approx 0.05$). To have 95% confidence that this is true for 99% of the talkers, the same test would have to be given to 299 talkers ($0.99^{299} \approx 0.05$). Similar calculations give the figures for the control application.

Because of the difference in voice quality expected by the two systems, they should be tested under different degrees of voice degradation. There is no reason to suppose that the system which is best under good conditions will also be robust against the different types of degradation caused by the public telephone system. Hence, the test for the catalogue entry system should be done over the telephone network, whereas the test for the control system could be done under good acoustic conditions.

For each case, the specified conditions provide the opportunity to define tests that generalize well to the actual application. In the catalogue entry system, the ASR device could be tested in simulation mode. It might be installed in parallel with a human listener who listened to the caller and keyed the requests into the catalogue order system as if the listener were the ASR device. The responses should be those that would occur if the ASR device were really doing the input. The error rates and dialect difficulties of the ASR device could then be monitored using the actual distributions of talker types to be expected in practice. The control system, on the other hand, could be tested by bringing the actual operators in for training with the ASR device, followed by extensive testing with each individual operator.

Research and Manufacturer-oriented Testing

The tests suggested for the catalogue and control systems were examples of "application-oriented" testing. Neither the manufacturers of the ASR devices nor the end-users can be expected to carry out the large-scale testing required for either application. For the end-user to perform these tests would cost far more than the cost of buying all competing systems and trying them out in parallel. The manufacturer would have to consider not only these two applications, but also all the other potential applications for which the machine might be useful. Hence, except for very large purchases or very critical applications with high costs of failure, these tests are unlikely to be done.

"Research-oriented testing" differs in scope and motivation from application-oriented testing. In developing an ASR device, the researcher need not be aware of all the potential applications, but will require information about the modes of failure for the device. Does it require well trained talkers? Does it fail on particular phonetic contrasts that are easy for human listeners? Does it fail with changes in speaking rate? There are many possible questions of this kind, and each ASR designer will have a different list, depending on the problems at hand. It is not possible for exhaustive tests to be made on the device under development, and neither is it necessary. The designer probably is aware of the potential weak points in the design, and can develop tests directed at those specific points.

One pitfall which may trap an ASR designer is the natural tendency to test using a few selected talkers or a pre-recorded speech database. This can lead to a design adapted to the idiosyncracies of those talkers, and not to the general user population. Nevertheless, the testing demands of a developmental system are clearly different from those required for comparative evaluation of a set of available ASR devices.

"Manufacturer-Oriented Testing" is aimed at providing a standardized set of tests. These tests should aid manufacturers to determine how their machines stand up to the competition, and aid potential users to select a small group of ASR devices as being competent for the application, without being over-designed and costly.

A gross measure of overall capability can serve to order the devices approximately. For this, we define an *Effective Vocabulary Capability* (EVC), on the assumption that the most important factor in many applications is whether the ASR device could handle the application vocabulary. Crudely stated, we define the EVC of a device as the size of the largest optimally chosen vocabulary that can be recognized with a tolerably small error rate.

We assert also that most application vocabularies can in principle be calibrated in difficulty, using a measure called *Required Vocabulary Capability* (RVC). The RVC may be defined as the size of an ideally-distributed vocabulary that would be as difficult as the actual vocabulary. It is numerically equal to the EVC of a device that would be just adequate to recognize the actual vocabulary with an acceptably small error rate. When the RVC for an application exceeds the EVC of a device, that device should be eliminated from consideration for the application. (The concept of RVC is due to R.K. Moore, personal communication, Oct. 1980). The degree by which the EVC of the device exceeds the RVC of the application is a rough measure of the security with which the device may be used in the application. Moore (1977) defined the concept of Human Equivalent Noise Response (HENR), which could be used to compare the difficulty of two vocabularies, and provided an algorithm for its computation. No detailed work has been done to covert this algorithm into an algorithm for computing RVC.

We do not specify here how the EVC of a machine or the RVC of an application can be determined. Research is under way into very rapid techniques for measuring EVC, in conjunction with the work of RSG-10. The method is based on computing the volume of a multidimensional space within which a vocabulary can be described, and comparing that volume with the volume associated with variability in the production of the "same" word on different occasions or by different talkers. Preliminary results show that useful measures of the EVC of a reasonably good machine can be determined with a few tens of utterances, rather than a few thousand.

Issues for Manufacturer-Oriented Testing

Even for isolated word recognizers, there are many independent considerations in evaluating ASR devices. Manufacturer-oriented testing cannot take into account the varying priorities of potential applications, but should give some information impartially about most of the more commonly important considerations. A partial list of aspects to be considered might include:

Performance considerations:

- Accuracy: How well does the device recognize a test vocabulary standardized for the class of ASR device, after a standard amount of training, with a standard set of talkers?
- Vocabulary Capability: How large a vocabulary can be used at one instant without excessively high error rates?
- Trainability of the machine to the talker: How easy is it to train the device to new talkers, and how likely is it that a new talker will achieve some rated accuracy? What proportion of the population can it be trained to recognize?
- Trainability of the talkers to the machine: How easy is it for a new talker to adapt to the requirements of the machine? How much improvement in performance can be obtained by deliberate training of the talkers?
- Talker independence: How capable is the machine of recognizing speech from a talker other than the ones for whom it has been trained?
- Language independence: Will the machine work equally well for talkers of different languages, in which the key phonetic contrasts may differ?
- Noise and distortion immunity: How susceptible is the machine to
  - — Additive, possibly varying, noise
  - — Non-linear distortion, including peak-clipping
  - — Phase distortion
  - — Amplitude variation
  - — Multi-path and other spectral distortion
- Speech mode independence: How tolerant is the machine of variations in speaking rate? Can it handle continuous speech, rapid staccato speech, or speech with irregular pauses between words?

Useability Considerations:

- Vocabulary flexibility: How easily can the user install a new vocabulary? Can the vocabulary be made to change as a function of what has just been recognized? If so, can the user install a new grammar to direct the changes of vocabulary?
- Sensitivity to types of grammar: Does the device depend heavily on redundancy in the task grammar? Is it affected by the presence or absence of easily recognized nodes in the grammar?
- Response speed: How much delay is there between the speaking of a word and the device's response? How many words per second can it respond to, independent of response delay?
- Ease of use: Does the machine require special setting up? Does it require microphones to be worn by the talkers, or can it be used over the public telephone network? Must the talker perform some introductory ritual every time he uses the device? Must the talker refrain from coughing or making other extraneous noises?

We will not consider conceptually trivial issues such as storage capacity, power consumption, physical size or physical ruggedness, important though such questions may be in a particular application. They can be specified in ways standard to electronic devices.

Although continuous speech recognizers are now commercially feasible, their evaluation poses a range of problems beyond those posed above. We shall mention continuous word recognizers at some points, but most of the following discussion should be understood as dealing primarily with isolated word recognizers.

Accuracy

Measuring the accuracy of an ASR device seems at first sight a straightforward problem answerable by standard statistical methods. In fact, the simplicity of the question covers a multitude of subtle problems. We will develop these problems through consideration of a simple example, which will be used as a basis for discussion in several of the following sections. Moore (1977) addressed the problem of evaluating the accuracy of ASR devices tested with vocabularies of varying difficulty. He used the concept of "Human Equivalent Noise Response" to describe the performance of an ASR device, and provided a formula that could be used to determine the difficulty of a vocabulary. We shall take a different approach to the performance measurement of ASR devices, but will build on Moore's approach to the determination of vocabulary difficulty.

A simple example

Consider a specific, relatively easy example. The vocabulary will be the digits 0–9, plus "start", "stop", "backup", "erase", and "end-group". All speech will be over a good quality microphone in a quiet environment. Only isolated words will be recognized. There are 15 words to be considered, but only 14 of these should be thought of as "vocabulary" words. The word "start" alerts the device to begin recognizing the test vocabulary. "Start" can therefore never be an appropriate response if any of the others are available. Conversely, "start" is to be recognized no matter what else is going on, such as conversation. It is a "keyword", and requires special treatment.

The test would be conducted separately for each talker. The talker would start by training the device, according to whatever methods were selected by the designer. Let us say that the device presents words from the vocabulary, and the talker repeats them. After sufficient repetitions, the device is said to be trained, and the test can begin. This statement begs the question of how the device responds to different amounts of training, and of whether each talker should be required to use the same amount of training regardless of the "intrinsic" difficulty of recognizing his/her voice.

In the test, the talkers attempt to enter prespecified words or numbers and the device responses are recorded. In the end, a confusion matrix is derived, which shows how often each possible response is made for each of the spoken vocabulary words. If the device is any good at all, most of this matrix will be empty, and most of the confusions will be between selected pairs of words, such as *five* for *nine*, or *three* for *two*. If all 14 vocabulary words were included, and, to take an extreme possibility, the confusion matrix were empty except for complete uncertainty between the pairs mentioned, the device would be recorded as achieving 86% correct. The error rate does not well represent the behaviour of the device, though, since the device makes 50% errors when the talker says 2, 3, 5, or 9. Another ASR device could give the same error rate by getting 86% correct, whatever word is spoken. For many applications, the former device would be preferable, since the vocabulary words on which the errors concentrate could be treated specially by the application designers. In the latter case, all input would be suspect.

One might expect the two ASR devices to differ in another important way. If they were used with larger recognition vocabularies, the device with distributed errors would probably show a marked increase in its error rate. On the other hand, the device whose errors were limited to specifically difficult pairs might be able to accept many new vocabulary words with no additional trouble. Its overall error rate might even decrease, if all the new vocabulary words were easy to discriminate, leaving only the original difficult pairs to contribute errors. The EVC of the latter device would be quite large, whereas the EVC of the device with distributed errors would be somewhat smaller than

the original test vocabulary.

Keywords vs. vocabulary words

Considering the same example, what do we do about the keyword "Start"? It really should not be considered together with the rest of the vocabulary, since the device should not be looking for "Start" at the same time as it is looking for the rest of the vocabulary. "Start" is possible only in context of a wide range of other conversational and non-conversational interference, and at that time, it is the only legitimate response. The rest of the vocabulary also occurs in the presence of interfering background, but then the device has to choose which word was spoken as well as whether the sound it heard was actually intended as a word. We will not make specific recommendations for testing keywords, and will rely on the hope that a device that is good at recognizing will also be good at extracting keywords from an ongoing background. In most cases, this will be a reasonable assumption.

Generalization

The EVC of an ASR device can be measured on one talker. A manufacturer wishing to make a good impression on the marketplace could select the talker best suited to the machine, and obtain a high EVC rating. We regard this as acceptable. An EVC rating obtained under these conditions is a maximum EVC for a given machine, and all manufacturers have an equal opportunity to optimize the talkers for their machines. EVC is in principle independent of the test vocabulary, so choice of vocabulary should not matter to the comparison. However, such an optimized EVC does not satisfy the potential purchaser's demand to know how the device will perform in practice. For this, one needs to know how well the result can be generalized.

Generalizing across talkers

There are recognized techniques for estimating how well measurements on a set of samples represent the "population value" of a parameter. The standard error of estimate varies inversely with the square root of the number of observations, and so forth. These techniques apply only under the assumption that the measured samples properly represent the population over which the measurement must be generalized. Unfortunately, in the real world, samples are seldom drawn from a homogeneous population. This is particularly true of talker populations.

The possible users of an ASR device can be related to any given talker by a series of ill-defined degrees of kinship. The closest relationship will be between a talker at one time and the same talker at another time. Next is probably family members having the same voice quality, then other same-sex family members, then other members of the same dialect group, then people of the same region, then talkers of

the same language, then foreigners speaking the same language, and lastly, foreigners speaking their own language. It is not at all clear where the sex difference in voice quality would fit into this series of levels. It should probably be considered independently, but in any case, results from a talker of one sex would not necessarily be expected to generalize to talkers of the other.

Talkers differ in voice quality (e.g. nasality or pitch) and speech behaviour (e.g. one might have a tendency to make irrelevant sounds). These differences are quite unrelated to dialect, language, or sex. Some of them might be important for a particular ASR device, some irrelevant. Their importance should be tested by using an adequate number of talkers sampled from each relationship level of interest. The result should be a curve showing the probability that the device will achieve any given EVC for talkers of that relationship level. Examples of such curves are shown in Fig. 1. A set of such curves for different relationship levels would be very valuable to potential users, some of whom might want a device with a high EVC for a few selected talkers, whereas others might want a device with a moderate EVC for most talkers from some group or other.

The curves need not be presented in detail to be useful. For example, suppose a device were developed and tuned by phoneticians at the University of Chicago. For such a machine, an adequate abstraction might be "80% of native English Speakers can attain an EVC of 130; 80% of white Mid-Western US natives can attain an EVC of 245; 80% of highly trained male phoneticians at University of Chicago gan attain an EVC of 460." If the application had an RVC of 350, this device could be used, but only for a fairly restricted and trained population of users. But if the application could be altered so that it had an RVC of only 80, the same device could probably be used for public access. Remember, the EVC does not depend on the test vocabulary, so that there could have been no optimizing of vocabularies for the different kinds of talkers.

The sampling problems associated with such measures are technically the same as those of public opinion polling. It is necessary to test with people representative of all the different voices expected to use the device. It would be substantially more time-consuming than a public-opinion poll, however, and probably more difficult to obtain talkers from many groups. Hence, it is probably impractical to demand that manufacturers perform EVC measures on a large number of diverse talkers. A properly sampled test might, however, be done on one or two representative devices, under the auspices of some national or international funding body. The results of these tests might give a guide to the extrapolation of results from a smaller population to a larger, more varied one. If so, manufacturers could be asked to provide EVC probability estimates for talkers sampled in a specified way from an accessible group, and a standardized technique might then be used to give rough estimates of the probability functions over the more

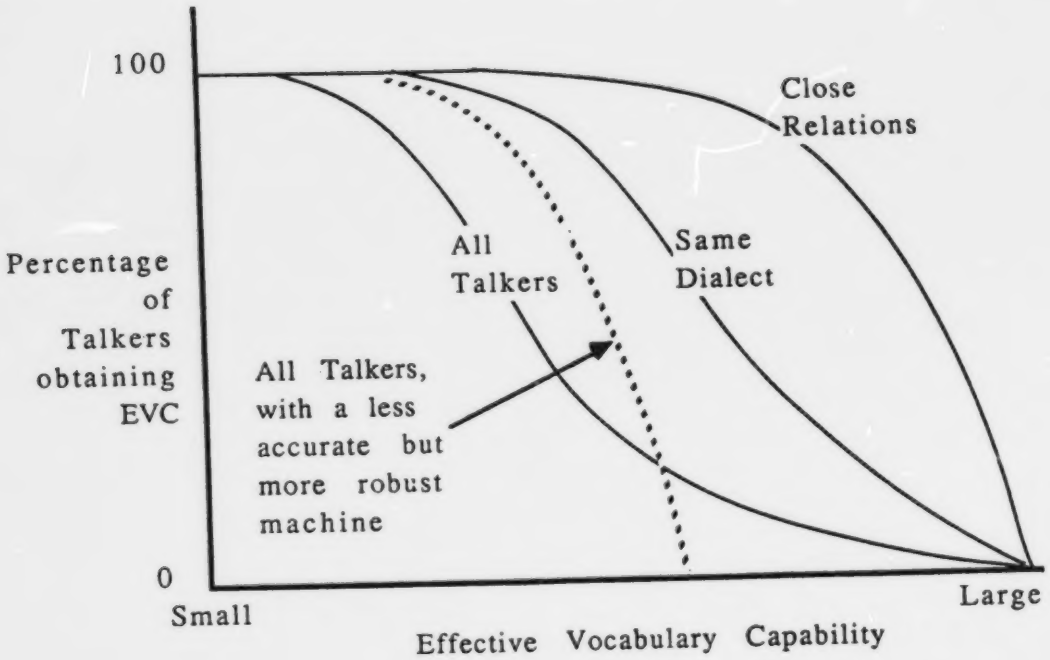


Figure 1 — Probability of achieving a given EVC

general population.

Generalization across languages

For military operations, especially within NATO, it is essential that a standard ASR system be able to deal with a variety of languages. Different languages rely on different phonetic contrasts. A distinction that is significant in English may be irrelevant in Greek, and vice-versa. An ASR device whose principles work well for English talkers may fail when used with Greek talkers, even when they are speaking English. In the absence of a detailed phonetic analysis of the languages and an equally detailed appreciation of the machine algorithms, it is not possible to tell how an ASR device will work when used in a new language environment. A manufacturer wishing international sales for his device should then perform an EVC probability measure over a variety of

languages, sampled from the area of intended sales. This is probably the ultimate test of generalizability for an ASR device. If its EVC probability curve is reasonably independent of language, it will be a good candidate for use in internationally standardized applications, such as NATO military operations.

Trainability

Trainability poses two independent questions. Both deal with how the machine performs for a particular talker after a period of use. One question is how well talkers can change their method of talking to suit the machine; the other is how well the machine can adapt to the speech of a particular talker. These two causes are different, but they combine to produce a common effect—improved performance. It may sometimes be necessary to separate them, because operator training can be affected by the psychological environment of the job. Separation is difficult, but may be possible, as we will show.

For most ASR devices there is an enrolment procedure for new talkers. Talkers repeat words so that the ASR device may learn their manners of talking. Some devices require one sample per word of less than the entire vocabulary, whereas others require many repetitions of every word in the vocabulary. In most cases, the amount of repetition has a strong influence on the performance of the device. Some ASR devices continually adapt to each talker during operation. Recognitions of high confidence are used to make small changes in the recognition parameters, so that, as a talker's manner of speaking drifts over time, the device maintains its ability to recognize the altered speech. Other devices have separate training and operating modes, so that updating the reference parameters requires an overt operation. Still other devices have only the capability to remake the recognition parameters for a talker from scratch.

What quantitative measure usefully describes the speed of the enrolment procedure and the subsequent adaptation of the device to the talker? Changes in the EVC will be appropriate when the rapid measurement techniques under development are completed, but as a stopgap other measures are required. One possible measure is the rate of decrease of equivocation per word spoken in the enrolment procedure. Equivocation is the difference between the information capacity of the vocabulary and the information transmitted through the ASR device. As the vocabulary increases, its information capacity increases as log N, but the actual information transmitted through a device that makes errors will asymptote at some finite value. If the equivocation is measured for a range of vocabularies of different sizes, it will be zero for small vocabularies having no difficult contrasts, but for sizes larger than a certain threshold, no vocabulary will have a zero equivocation. As the vocabulary increases, the equivocation will approach log N minus a constant. The vocabulary size above which the equivocation starts to

rise appreciably could be taken as a measure of the EVC.

In early training, the equivocation for a moderate vocabulary presumably is high. During training it reduces at some rate that might perhaps be approximated by a declining exponential function asymptotic at a finite positive value rather than zero. Devices that require periodic training or re-enrolment sessions may also show improvement in performance over time, but this will be primarily because the talker learns to adapt to the machine. Nevertheless, a discrete version of the exponential model might be applicable as a first approximation to their performance changes.

For a given ASR device, it should be possible to relate the EVC rating for a talker to the equivocation for a moderate vocabulary after training is completed. One does not expect to be able to measure the equivocation for a range of vocabularies during training, and some kind of functional relationship must be assumed or found from experiment. It should then be possible to extrapolate this relationship to the earlier training stages, and report the equivocation in terms of an "equivalent EVC". The training requirements of the device could then be reported in terms of the initial EVC, the growth rate of the EVC, and time to reach a reasonably stable EVC. These values will, of course, be distributed over the various heterogeneous talker populations, and would be better reported in terms of cumulative probabilities of achieving a given level.

Training versus operational performance

The question of trainability and talker adaptation shades directly into the question of how well laboratory tests can relate to the operational situation. Apart from the usual issues of generalization—how well the test conditions simulate the operational conditions—transferring an ASR device into an operational situation immediately brings up the question of training users to the machine.

It is noticeable that people accustomed to talking to ASR devices have a tendency to speak clearly as a matter of habit; most of the population does not. Hence, very often a newly installed ASR device will show performance substantially worse than expected from laboratory studies. However, as time goes on, some users learn what is required of them to improve the performance of the machine, regardless of any adaptation the device may or may not be doing. It may be the case that some ASR devices encourage this kind of user training, whereas others inhibit it.

It should in principle be possible to devise some kind of index for the tendency of an ASR device to induce user training. The major problem might be that it is impossible in an operational context to distinguish user training from device adaptation. Even keeping speech samples from early training days for use in later tests would not work. Most reasonable devices would adapt as the talker's speech habits drifted, regardless of whether the drift of speech habits were favourable

or unfavourable to the device. Changes in performance on the recorded speech would be due to adapation by the device, but it would be impossible to tell whether any performance improvement on live speech were due to this adaptation or to talker training.

Long-term perfomance improvement is hard to measure in a laboratory. The motivation of laboratory talkers is different from that of operational users, and their results might well not be generalizable to the field. Perhaps the most effective way of measuring operational performance improvement is just to monitor the use of a device in the field over a period of time. It can hardly be called an "experiment", and cannot provide a measure for a new ASR device under consideration, but over time, a useful index for a popular device might be obtained.

A possible method of separating machine adaptation from operator training suggests itself, although it is not totally satisfactory. It might be possible, using field studies, to note the spread of performance improvement rates, or variation in the timing of sharp improvements in performance, thereby identifying operators who differ in the way their speech changes over time. This might permit analysis of periods over which most of the performance improvement was due to operator training as opposed to periods in which machine adaptation predominated.

Talker Independence

We have discussed the measurement of how well new talkers can be introduced to an ASR system, both in terms of the distribution of performance across talkers after training, and in terms of the time course of training. "Talker independence" means the ability of an ASR device to understand talkers to whom it is not trained. It therefore can be taken as a special case of the measurements we have already described. The index of talker independence of an ASR system is simply the measure of generalizability across talkers, taken before training.

It is unlikely that effective "talker-independent" ASR devices will be made without some adaptive capability. This means that as soon as the new talker has said a few words to the machine, it will understand him better than at first. This is as it should be, but it confuses the measurement problem. In testing an ASR system for "talker independence", its adaptation algorithm should be suppressed, so that new talkers can provide an adequate sample of speech to permit measurement of the response surface parameters. In use with a public access system, a performance measurement taken without adaptation is probably more meaningful than one permitting adaptation, since the typical interaction with a user requires the ASR system to make a reasonable response within the first few words. A "talker independent" system for a less public use might be permitted a little adaptation time for a new talker. Accordingly, a measure of the initial rapidity of training would be useful for such systems. Determining this measure is

not difficult in principle, but it does require either the specification of some standardized training procedure, or measurement in some realistically generalizable training environment. Either way, adaptation would have to be turned off after selected training increments, in order to define the device performance ratings at those stages.

The important consideration for purposes of testing is that there is no intrinsic difference between testing "talker independent" systems and "talker dependent" systems. The real difference is that the "talker independent" systems are tested very early in training, and measurement of the initial slope of the training performance gain is important. For "talker dependent" systems, the important measurements are those taken after training has approached asymptotic performance. Otherwise, the measures required are the same.

Training and the measurement of talker independence

At any stage of training, it is possible to define a set of talker population probability curves for the EVC (or equivalent EVC derived from sub-vocabulary equivocation). These curves indicate the proportion of the sampled population achieving an EVC of at least the indicated amount. As training procedes, the proportion of the population achieving a given EVC should increase. If talker proportion is plotted against EVC and time, it forms a surface whose properties reflect both the trainability of the machine and its talker independence, where talker independence refers both to its ability to understand talkers to whom it is not trained and to its ability to be trained to a range of new talkers.

The shape of the performance surface is unknown, and must be determined experimentally for a variety of ASR devices. With luck, different devices will define a family of surfaces that could be described in terms of their behaviour on each axis separately. This is to say that it is conceivable that a transformation of the EVC might result in a surface describable as $F(EVC) \approx T(\text{training}) \phi P(\text{proportion of population})$, where ϕ represents some arbitrary operator. If so, then describing T and P separately would describe the whole performance of the ASR device in respect of training and talker independence. These could then be used to determine the relative suitability of different devices for public access (high initial P, does not matter what T looks like) or for a small group of highly trained and selected users (high values for $P(0)$ or $T(\infty)$). Manufacturers could present curves of P(population proportion) and T(training), for different population sampling ranges.

Noise and Distortion Immunity

Quantification of the noise sensitivity of an ASR device will be rather difficult in the general case. "Noise" refers generally to erratic and unpredictable events. It can be an extraneous signal added to the desired signal (additive noise), such as impulse noise, tonal noise, or broadband noise; it can also refer to erratic variations in the transmission of the signal (multiplicative noise), which can include dynamic spectral variations such as may be caused by multi-path fading in radio transmission. More complex forms of noise are possible as well. They all have the common feature that their effects are either unpredictable in detail or unrelated to the desired signal. "Distortion", on the other hand, refers to predictable modification of the signal. and includes spectral filtering, non-linearities in transmission (such as peak-clipping), phase errors, and so forth.

Different ASR devices may have vastly different sensitivities to the various kinds of noise and distortion. Accordingly, a manufacturer-oriented testing program must account for at least a reasonable range of types of noise and distortion. However, even testing an ASR device in the clear is a laborious task. Testing the sensitivity of the various parameters to different varieties of noise and distortion multiplies the task by a substantial factor.

The noise amounts are not easy to measure. How does one equate the filtering in a telephone connection with the multi-path fading of a radio link, with impulse noise, with white gaussian noise, with competing speech, etc? These different kinds of noises may well have different effects on different kinds of ASR devices, and it is really only in connection with their masking effects that noises can be effectively measured. We have a circular definitional problem: the quantification of how much noise is needed to have a given detrimental effect on an ASR device is determined in part by how the particular kind of noise, as compared to other kinds of noise, affects ASR devices in general. There is no obvious way of making a correspondence between noises other than by referencing the task over which they act as noise. It may well prove necessary to challenge the ASR device with a set of standardized noises and distortions, using both good (large EVC) and poor talkers to assess the effects.

Speech mode independence

Isolated-word recognition devices are designed to work with a variety of speech modes. Some require words to be separated by 200 msec. of silence, others require only a short but definite pause which gives a staccato quality to the speech, while others can handle naturally spoken phrases. The range of tolerance of a device for different speech modes is an important parameter of its performance. There are many variants of the concept of "speech mode". Here, we will concentrate on speech rate, which can be divided into word duration and inter-word

gap duration.

Most ASR devices are optimized for some particular speech mode, including pause duration. Frequently, the modes can be characterized by a conventional term, such as "deliberate", "staccato", or "natural". It may be possible to check the effect of gap duration by recording the speech necessary for a "deliberate" test, and then reducing the inter-word gaps on the recording. However, most isolated word recognizers have an adjustable threshold gap to discriminate between intra-word pauses and inter-word pauses, and another threshold to distinguish inter-word pauses from termination of a speech group. A test of the effect of inter-word gap would not be very useful for such a device. It might be useful for a continuous speech recognizer, since inter-word and intra-word gaps are recognized as part of the recognition process, not as an initial segmentation operation. However, for continuous word recognizers, an isolated word speech mode is unlikely to be used very often, and the result of the test might not be worth the effort invested in it.

For practical purposes, precise control of word duration and inter-word gap is not required. Comparative evaluation of devices does require that some common test parameters be employed, however. It is perhaps here, more than in any other facet of the evaluation of ASR devices, that a common test recording might have value. If the basic performance capability of the device were determined by live testing according to standard protocols, then the additional use of a standard test recording for assessment of speech mode flexibility would be possible. Any tuning of performance for the particular recording would show up on all speech modes, but hopefully not on the relative effect of different speech modes.

A recording for testing speech mode should contain test material spoken by a few talkers of a range of dialects and of both sexes, with accompanying training material for each talker. Each talker should speak the test material in three or four modes, ranging from very deliberate to continuous. The test material itself should contain both difficult pairs and a full range of phonetic contrasts, but need not use a large vocabulary.

Useability Considerations

Thus far, we have dealt with questions that can, in principle, be answered numerically. In what follows, the questions are as important for assessing the probable value of an ASR device for a particular application, but their answers are less quantifiable. In some respects, it is hard to obtain more than a subjective judgement, which is useful for an advertising executive, but not for a potential purchaser.

Vocabulary Flexibility

At any moment in its operation, an ASR device is attempting to recognize a spoken word as belonging to a currently active vocabulary. The ease of changing this vocabulary varies from machine to machine. Some can change only if the user deliberately reads in or trains a new set of templates, whereas others can move flexibly from one active vocabulary to another according to the recognized grammatical context. The flexibility of vocabulary change may or may not be important in a given application.

A trade-off between large EVC and good vocabulary flexibility is sometimes possible. To take an extreme example, let us suppose that the application requires a vocabulary having an RVC of 300, but that at any point in the recognition process only a restricted sub-vocabulary with a maximum RVC of 30 is needed. A device with an EVC greater than 300 could handle this task with no help from syntax, but a device with an EVC of only 30 could be used if it could change its active vocabulary after each word recognized.

If the EVC for the device were substantially more than 30, recognition should be essentially error-free. After each word recognized, only those words legally following it in the grammar would need to be included in the sub-vocabulary for the next recognition. If, however, there were a reasonable probability of error, then those words following each of the likely targets would have to be scanned. There would be several sub-vocabularies, each having an RVC of up to 30. If these were to be combined into a single recognition scan, they would form a larger sub-vocabulary with an unpredictable RVC. Regardless of the actual increase in RVC, which could easily be more than the sum of the component RVCs, the EVC for the device would then have to be greater than 30 to accomodate the increase. But having a device with an EVC substantially greater than 30 would reduce the error rate on the individual words, thus reducing the combinatorial possibilities for subsequent recognitions. Hence, performance will be a steeper function of EVC for devices using task grammars than for devices with a fixed vocabulary.

If the application has a task grammar which permits active subvocabularies having individual RVCs much less than the RVC of the total vocabulary, then the EVC of an acceptable device must be substantially more than the RVC for the subvocabularies. This contrasts with the situation for tasks with fixed vocabularies, in which the EVC for the device simply needs to exceed the RVC of the vocabulary by some standard safety margin. For grammar-controlled applications, the safety margin must be much larger. In practice, however, there will usually be only a few points in the grammar at which a large RVC is required, and these will control the EVC requirement for the device.

Sensitivity to types of grammar

The foregoing leads directly to the question of how sensitive the ASR device performance will be to the type of grammar used in the application. If it is possible to define nodes in the grammar so that what follows the node is independent of what came before, the grammar is piecewise non-redundant. If, however, the earlier words influence the possibilities for later words and vice-versa, the grammar is redundant.

As an example of a redundant grammar, consider the following, which gives rise to four reasonably likely sentences.

$$\text{The} \begin{bmatrix}\text{plane}\\ \text{car}\\ \text{book}\\ \text{tide}\end{bmatrix} \text{was on the} \begin{bmatrix}\text{runway}\\ \text{road}\\ \text{shelf}\\ \text{ebb}\end{bmatrix}$$

Without the semantic links (which we assume to be formalized into an acceptable grammar), 16 sentences would be possible. Using the redundancy, the ASR device could completely miss a word out of either the first or the second group, while getting the sentence completely correct. More effectively, it could use the contingencies among the words to reinforce the probabilities of its selections in both groups. As an alternative, in which the possibilities following a node are independent of what went before, consider

$$\begin{bmatrix}\text{Two}\\ \text{Three}\\ \text{Four}\end{bmatrix} \begin{bmatrix}\text{buses}\\ \text{cars}\\ \text{vans}\end{bmatrix} \text{passed the observer.}$$

The implications of the different types of grammar on the EVC for the device are important. If the device can use redundancy in the grammar, then the EVC will be apparently increased for those subvocabularies whose interpretation can be aided by other parts of the message. If not, it may be apparently reduced by the mechanism described in the preceding section. It would be possible, then, to use a vocabulary whose RVC was larger than the EVC of the candidate ASR system, provided that those sections of the grammar yielding the large RVC could be linked with redundant words elsewhere in the permitted utterances.

In evaluating an ASR device, then, a potential purchaser might need to know whether it can accept a grammar description which will permit designation of redundancies, and whether it can take advantage of those redundancies.

Response Speed

Just as with some of the other aspects of testing ASR devices, testing response speed seems at first trivial, then becomes more subtle on investigation. The term "response speed" could refer to

- Response delay: The time after a word is spoken until its recognition response is available
- Response rate: The highest rate in words/sec for which the ASR device can maintain its performance
- Throughput: The highest rate at which a talker can input essentially correct data, allowing for feedback and correction (a system performance measure)

It is very likely that they all are affected by the size of the attempted vocabulary as compared to the effective vocabulary size of the device. In addition, throughput rate will be affected by accuracy, because of human factors discussed below.

Response delay

Response delay is important in situations where the user expects to use feedback for correcting the ASR device's responses. The response delay to an utterance is the important quantity. For an isolated word recognizer, this may be the word, whereas for the continuous speech recognizer, it is an indeterminate quantity, perhaps a phrase, perhaps a complete query or instruction. In the course of other evaluation testing, the response delay can be measured for each sample of each word spoken. It could be measured from the start of the spoken word or from its end. There are arguments for either mode of measurement. In favour of measuring from the beginning of the word is the fact that information about the word is being presented to the ASR device continuously from the beginning of its sound pattern. Hence, a search process could be reducing the possibilities as the word is spoken. In favour of starting the measure at the end of the word is the fact that words come in different lengths, and for some of them it is impossible to be sure of the identity of the word until it is finished. In operational use, the delay from the end of the utterance is probably more important than the delay from its start. We recommend that response delay distributions be measured from the ends of utterances.

Using a human as an ASR device, accuracy can be traded for speed. If the human is forced to respond very rapidly, the resulting accuracy is lowered. This "speed-accuracy trade-off" is well documented for many decision problems in psychology. Many artificial ASR devices have algorithmically determined response delays, and cannot be speeded. The possibility exists, however, that more sophisticated machines may be able to trade speed and accuracy. For example, when dealing with grammatical nodes having a small possible vocabulary, template matching time may be drastically reduced without compromising accuracy,

thus permitting more time to be spent on words that are difficult to recognize. The capability of a machine to make a speed-accuracy trade-off should be measurable. A related measure would be the relationship of response delay to attempted vocabulary size and accuracy. For testing purposes, the response delay measurements could be integrated with the measure of effective vocabulary size. Since the response delay may or may not vary with the size of the testing vocabulary, it should be measured at several vocabulary sizes.

The use of redundancy in the grammar poses a problem in the measurement of response delay. If the device has not made a high-probability identification of some word, and the grammar suggests that a redundant subsequent word is coming, the device should delay interpretation until the subsequent word occurs. Such a delay might sometimes be disconcerting or even incapacitating to an operator who might be speaking and monitoring well-isolated words (Taylor, 1986). On the other hand, redundant grammars may not be likely to occur in situations where monitoring the feedback makes sense for individual words.

Response rate

In the most simple-minded devices, a new word cannot be processed until recognition is completed on the old word. For such a device, response rate will just be the inverse of response delay. More sophisticated devices can use a pipe-line processing technique, in which early stages of processing are being done on the new word while final processing is still continuing on the old. Such overlapped operations permit response rates much faster than the inverse of the response delay. The effect is that the response to one word happens during or even after the presentation of the next. In continuous speech recognizers, recognition of one word may require presentation of later words. For example, in the beam-search technique used in the HARPY project (e.g. Klatt, 1977), words on one path are not treated as recognized until the whole path has been recognized or until all competing paths containing alternate interpretations for that chunk of sound have been eliminated. The response delay may be very large and erratic for early words in utterances, but the response rate can still be uniform and high.

Throughput rate

The throughput rate is the most interesting of all for a potential user. It is not a characteristic of the device, since it depends on methods of feedback just as much as on the ASR response rate. It also depends on recognition accuracy in a peculiar way (Lancé, personal communication, May 1979). For error rates above a certain threshold, users tend to check for errors and correct them, obtaining slow but accurate performance. For smaller error rates, they tend to assume, perhaps falsely, that the response will be correct, and therefore do not closely monitor the available feedback. This behaviour yields fast, but

less accurate performance. The function relating transmitted information rate to accuracy then would have a break point at the threshold accuracy. It is not *a priori* obvious whether the function would be higher or lower for ASR device accuracies above the threshold.

It is likely that the threshold accuracy is related to the probability per word of error, not to the information transmitted per word. Hence, increasing the test vocabulary size might increase the information per word, but decrease the probability of a word being correctly recognized. At some point, the decrease would give accuracies below the threshold and thus probably decrease the information rate. Again, questions of feedback and correction technique might control whether throughput was higher above or below the threshold.

For purposes of measuring the information throughput capability of the ASR system, the existence of the threshold accuracy poses a significant problem. It will undoubtedly shift with operator training and with the perceived importance of input accuracy as opposed to input speed. Hence no absolute measure can be derived. There should, however, be some reasonable way of ranking ASR systems in respect of maximum information throughput by trained operators. Possibly this kind of measure should be considered application-dependent rather than being included in a suite of manufacturer-oriented tests.

Ease of Use

The ease of use of an ASR system depends only in part on the actual ASR device. As in the case of the throughput rate, with which it is related, ease of use depends heavily on the rest of the application system. There are one or two factors, however, which do depend on the device.

Ease of use is a subjective quality, not readily measured in any abstract sense. A head mounted microphone may be easier to use than a telephone handset in many applications (e.g. maintenance, inspection, etc.), but in others it would be intolerably difficult (public access catalogue ordering). About the best that can be done, from the manufacturer's point of view, is to specify the requirements for usage. These specifications might include the need (or otherwise) for noise-cancelling microphones, quiet environments, and rigidly controlled talking techniques.

For many ASR devices, it is probable that the need for rigorous conditions depends on the application. Applications with smaller vocabulary requirements may be able to get away with some relaxation of conditions. If so, then a method of specification might be to give the EVC with (say) head-mounted noise-cancelling microphones and with telephones over the public network. The application user should be able to determine the RVC, and thus to determine whether the device would be adequate given the application constraints on microphone type.

Testing Issues

Standardized vs. Randomized or Application Vocabularies

For applications-oriented testing, the vocabulary used to measure the performance of an ASR device should ideally be related to the application. This approach would test not only the basic performance of the device as measured by the (vocabulary-independent) EVC, but also any quirks that might cause errors specific to the application vocabulary. In manufacturer-oriented testing, however, the quirks cannot be specifically tested. There are too many possibilities, and the results are supposed to be used in evaluating the device for a wide and unknown range of applications. It might, therefore, be valuable to test the device against a standardized vocabulary that challenges those aspects of ASR devices most likely to cause trouble.

One could take two possible approaches to the design of a standardized test vocabulary: (i) use a wide-ranging vocabulary and try to assess the maximum contingent uncertainty achievable by a device, or (ii) use a vocabulary stressing difficult contrasts, to maximize the ability of the test to force errors. Both approaches have disadvantages as well as advantages. To use a wide-ranging vocabulary is to require extensive training and testing, and is thus very costly. To use a difficult-contrast vocabulary is to run the risk of missing characteristic errors which might be made by the device, or of over-emphasising other kinds of difficulties. Word pairs which are easily discriminated by humans may not be easy for a particular ASR device, and vice-versa. A modified minimum-contrast vocabulary would therefore seem to be an appropriate test.

Another approach would be to use a randomized vocabulary, with words selected by an appropriate sampling method from a dictionary. This technique has the advantage that phonetic contrasts that might unexpectedly provide difficulty to the ASR system have a good chance to be selected. In a deliberate vocabulary selection, emphasis is likely to be placed on contrasts expected to be difficult. The disadvantage of the method is that scattered results will be obtained from different vocabulary samples of a common size.

Recorded vs live test speech

There has been considerable debate about the construction of recorded speech data bases, and whether they are even desireable. Some argue that they would provide a standardized reference for performance claims; others point out that different ASR devices prefer different speaking modes (staccato, minimal inter-word pauses, continuous speech, long inter-word pauses, etc.) and that a test recording suiting one device might not suit an equally competent device which

preferred a different mode. The range of tolerance of speech modes may well be an important factor in evaluating an ASR device. It poses a significant problem, however, in designing recorded speech data bases that are fair to all ASR devices. A further difficulty arises from the differences in scores obtained when testing a speech recognition device and when using it operationally.

The problem is not intrinsic, in the sense that a data base could in principle include a range of speech modes, from continuous to well isolated words. Making such a data base is several times more time-consuming and expensive than making a data base in one speech mode. One of the critical areas of difficulty in constructing standardized tests of ASR devices is that the required amount of speech becomes very large. Adding new dimensions of testing multiplies the required amount of speech by an appreciable factor.

Another criticism of standard recorded test data bases is that it would be possible for manufacturers to tune their devices for good performance on just those speech samples. In principle, it would be possible to recognize perfectly every utterance on the recording, if the recording were available for training purposes before the test; and such a public data-base test recording would have to be made available. No published results based on such recordings could be relied on. Even if specific tuning for the recording were not done, the machine could and probably would be developed with algorithms and parameters known to get good results on the standard recordings. This is not what is wanted in a reference standard test procedure. A well funded body might periodically make new standard test recordings, but this seems to be an unnecessary and expensive operation.

Standardized testing of ASR devices should refer to a procedure, not to the actual speech samples used in the tests. A standard vocabulary or set of vocabularies could be defined; a standard way of sampling and training talkers, standards for noise-free and noisy conditions, and so forth could be specified. Tests performed using these standardized conditions should be reasonably commensurate.

No matter what the technique, testing ASR devices under a wide range of conditions will require the gathering of a large amount of data. In experimental psychophysics, problems of excessive data requirements are handled in two different ways: separating the dimensions of the problem, and using efficient adaptive techniques. In testing ASR devices, it may be possible to separate dimensions. For example, a machine may have the same characteristic failure on irregularly timed words regardless of the accent of the talker. One could separately test the effect of speech mode using one or two talkers, and the talker independence using several talkers in the optimum mode.

A fully live testing program presents its own problems. A representative set of talkers untrained in speaking to ASR devices must be obtained. The talkers should represent the range of dialects and speech habits of the population the device is expected to deal with, and should

be capable of accurately speaking the test words. Selection of a group of test talkers seems to be virtually impossible, because after the talkers have been used in a test they will be trained in talking to ASR devices, and will thus be ineligible for other tests.

If both testing based on recordings and live testing are unsatisfactory, what remains? It seems likely that the testing program should include both recording and live phases, and should be done progressively. For example, an initial recording-based phase might assess the vocabulary size for a few talkers using one speech mode. The speech mode might be selected from a library of three—well isolated words, staccato words, and continuous speech. After this, a "challenge" vocabulary of ordinarily difficult discriminations might be used with many different styles of speaking, to determine the range of talkers, and a final phase might use live talking over a long period with one or two good talkers, to assess the ultimate capability of the device. The recording base is used in this paradigm to set a rough level of capability for the device, to aid selection of an appropriate vocabulary size for EVC measurement in the live situation. Only measures taken in the live situation would be used for public claims.

Conclusions and Recommendations

Effective Vocabulary Capability

Comparative evaluation of isolated word ASR devices is possible. The key concept is the reduction of performance data to a measure independent of the vocabulary used for testing. This paper has introduced the concept of *Effective Vocabulary Capability*, which is an estimate of the largest possible vocabulary the device could handle with an acceptably low error rate. Unlike percentage correct, EVC is independent of the test vocabulary and is necessarily monotonic with recognizer capability.

The EVC may be computed directly, by challenging the ASR device with a variety of vocabularies of different sizes. For this computation, the ASR device is treated as an information channel; the test vocabulary is the input, and the responses the output. The equivocation between input and output determines the error level. This measure is less sensitive to difficult pairs of vocabulary words than is the simple error probability. The EVC is taken to be the vocabulary size at which the equivocation begins to increase sharply.

The direct measure of EVC takes a large amount of speech data for each determination. Not only must the device be trained to each talker, but also there must be enough test samples per vocabulary word to allow a reasonable measure of the probability of various responses to that word. Furthermore, the procedure must be repeated for several new vocabularies of different sizes. Work is in progress on much

quicker ways of estimating approximate values for the EVC of a recognition device, using tens, rather than thousands of utterances.

The EVC provides a simple monotonic measure that can serve to rank ASR devices in order of gross capability. It does not give any indication about specific difficulties a device may have, that might be relevant in a particular application.

Required Vocabulary Capability for an Application Vocabulary

Just as the EVC for a device allows the comparison of devices against a common scale, so the *Required Vocabulary Capability* for a vocabulary permits comparison of vocabulary difficulty, and determination of whether the vocabulary is too difficult for an ASR device with a given EVC. Conceptually, the RVC is the size of a vocabulary optimally designed to have no especially difficult word pairs, that would give the same overall performance as does the vocabulary being measured. The RVC cannot be smaller than the test vocabulary itself, and may be much greater, depending as it does on the most difficult contrasts in the test vocabulary. Because different ASR devices may be expected to have idiosyncratic confusion problems, the RVC is defined with reference to a human listener. Ideally, the RVC for an application vocabulary should be determined for each language and dialect group, but in practice a less accurate approximation should suffice.

Moore (1977) has provided a formula for assessing the confusions to be expected among words of a vocabulary in English, and the resulting matrix can in principle be used to determine a vocabulary space. Practical algorithms for doing this have not yet been studied. The expected confusion distances within this vocabulary space determine the size of the idealized vocabulary having a similar overall degree of difficulty, as measured by the equivocation in word recognition. This size is the RVC for the vocabulary. Moore's formula is presumably most valid for British English, but probably is good enough for other varieties of English as well. For other languages, and divergent forms of English, similar formulae could be obtained.

If greater precision were required for a vocabulary, possibly taking into account environmental noise conditions to be expected in operation of the ASR device, the application vocabulary could be directly tested using human listeners. The resulting confusion matrix could be used in place of Moore's derived one, and the same computations of RVC could be made.

A device can be tested under a wide range of conditions only if a fast measuring technique is available. As described in this report, EVC measurement is not fast. It is hoped, however, that work in progress will lead to EVC measures that can be performed with a few tens of utterances, thereby making it possible to implement the suggestions in this report.

Note

This paper is DCIEM Report 86-P-16

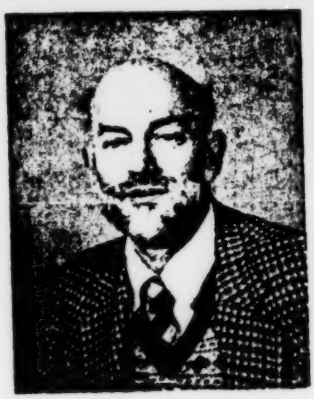
References

Kaplan, G. Words into action: I. *IEEE Spectrum*, 1980, *17*, 22–26.

Klatt, D.H. Review of the ARPA Speech Understanding Project. *J. acous. Soc. Amer.*, 1977, *62*, 1345–1366 (U).

Moore, R.K. Evaluating Speech Recognizers. *IEEE-Acoustics, Speech and Signal Processing*, 1977, *25*, 178–183.

Taylor, M. M. Layered Protocols in Voice Interaction with Computers. Proceedings of the AGARD AeroMedical Panel Symposium on *Information and Decision Making in Advanced Airborne Weapon Systems*, Toronto, April 1986.



Martin Taylor received his Ph.D. in Experimental Psychology in 1960 from The Johns Hopkins University following earlier degrees as an engineer at the University of Toronto and at Johns Hopkins. He has been involved with both research in perception and the development of concepts for computer systems for the Defence Research Medical Laboratories and its successor, the Defence and Civil Institute of Environmental Medicine, since 1960. At present, he is mainly concerned with the principles underlying computer-human interaction, and is a member of a NATO Research Study Group on Automatic Voice Processing. He was co-author, with his wife, of the book *The Psychology of Reading* (Academic Press, 1983), and has published some 60 refereed papers in various areas of psychology and computation.